Dilated Transformation-Guided Unsupervised Multimodal Learning for Hyperspectral and Multispectral Image Fusion

Yuanchao Su, *Senior Member, IEEE*, Sheng Li, Yicong Zhou, *Senior Member, IEEE*, Lianru Gao, *Senior Member, IEEE*, Mengying Jiang, Xu Sun, *Member, IEEE*, Haiwei Li, and Enke Hou

***Abstract*—Multimodal fusion widely uses convolutional layers to capture local correlations and adjust feature dimensions. However, the progressive expansion of the receptive field in convolutional layers often compromises spatial context retention, leading to the loss of fine details. Furthermore, the fixed-size kernels typically used in standard convolution restrict the network's ability to capture multiscale contextual details. To address this limitation, this article develops a dilated transformation-guided unsupervised multimodal learning (DTUML) method to fuse a high-resolution multispectral image (HR-MSI) and a low-resolution hyperspectral image (LR-HSI), thereby generating a high-resolution hyperspectral image (HR-HSI). Our DTUML adopts a dual-stream encoder architecture to conduct multimodal data, where one stream focuses on preserving spectral information from LR-HSIs, while the other emphasizes the acquisition of spatial details from HR-MSIs. These complementary features are subsequently integrated to ensure spectral fidelity and retain spatial detail. Then, a convolutional layer restores dimensional consistency and outputs an HR-HSI. Extensive experiments demonstrate the effectiveness of DTUML, showing superior performance and strong competitiveness compared to state-of-the-art methods. The code is available at https://github.com/yuanchaosu/TGRS-DTUML**

***Index Terms*—Dilated convolution, hyperspectral and multispectral image (MSI) fusion, hyperspectral image super-resolution, multimodal fusion, multimodal learning.**

I. Introduction

HYPERSPECTRAL images (HSIs) are renowned for their exceptional spectral resolution, capturing detailed spectral information across numerous bands [1], [2], [3], [4]. However, this strength often comes at the expense of spatial resolution, which is typically lower and limits the effective utilization of the data in various applications [5], [6], [7], [8]. Compared to HSIs, multispectral images (MSIs) are more focused on retaining spatial details and are less expensive to capture owing to the lower cost of multispectral sensors [9], [10]. In recent years, MSI–HSI image fusion has emerged as an important technique for enhancing the spatial resolution of hyperspectral imagery [11]. This technology enables hyperspectral image super-resolution by data fusion, effectively leveraging the complementary information of both modalities [12], [13], [14], [15]. Nowadays, HSI–MSI fusion can be categorized into two types: supervised and unsupervised approaches, depending on whether training samples need to be provided [16], [17], [18]. Supervised fusion methods need labeled training data to learn a mapping from low-resolution HSIs (LR-HSIs) and high-resolution MSIs (HR-MSIs) to output high-resolution HSIs (HR-HSIs) [9]. These methods usually require paired training samples and ground truth (GT) for learning rich spectral information from LR-HSI and the fine spatial details from HR-MSIs [19], [20]. In contrast, unsupervised fusion methods do not rely on labeled samples, particularly useful in real scenarios where GTs are unavailable or difficult to obtain [21]. Moreover, unsupervised methods usually need to simulate physical imaging to produce HR-HSIs, which means that the data fusion typically leverages spectral unmixing models to couple and constrain the generation of HR-HSIs [17], [18], [22], [23]. Traditional unsupervised fusion methods rely on the theories of statistics and machine learning, such as sparse representation, nonneg-

Received 8 November 2025; accepted 17 November 2025. Date of publication 1 December 2025; date of current version 11 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42571463; in part by Macau Young Scholars Program under Grant AM2023033; in part by Shaanxi Province Youth Science and Technology Star Program under Grant 2024ZC-KJXX-115; in part by the Natural Science Foundation Research Program of Shaanxi Province under Grant 2025JC-YBMS-257; in part by the Open Research Fund of Shaanxi Key Laboratory of Optical Remote Sensing and Intelligent Information Processing under Grant KF20250302; in part by the Science and Technology Development Fund, Macau, SAR, under Grant 0050/2024/AGJ; and in part by the University of Macau Development Foundation under Grant MYRG-GRG2024-00181-FST-UMDF. *(Corresponding author: Yicong Zhou.)*

Yuanchao Su is with the Department of Computer and Information Science, University of Macau, Macau, China, also with Shaanxi Key Laboratory of Optical Remote Sensing and Intelligent Information Processing, Xi'an 710119, China, also with the College of Geomatics, Xi'an University of Science and Technology, Xi'an 710054, China, and also with Shaanxi Yellow River Planning and Design Company Ltd., Yulin 719000, China (e-mail: suych3@xust.edu.cn).

Sheng Li is with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: shengli202302@126.com).

Yicong Zhou and Mengying Jiang are with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo; mengyingjianggdut@foxmail.com).

Lianru Gao and Xu Sun are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: gaolr@aircas.ac.cn; sunxu@aircas.ac.cn).

Haiwei Li is with Shaanxi Key Laboratory of Optical Remote Sensing and Intelligent Information Processing, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: lihaiwei@opt.ac.cn).

Enke Hou is with the College of Geology and Environment, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: houek@xust.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3636047

ative matrix (or tensor) factorization, and Bayesian inference [22], [23], [24].

Recently, advancements in deep learning (DL) have significantly advanced unsupervised HSI–MSI fusion, offering powerful frameworks for automatically extracting and integrating complex spatial and spectral features from both modalities [16], [25]. A wide range of novel DL-based fusion architectures has emerged in recent years. These include unsupervised convolutional neural networks (CNNs), generative adversarial networks, and deep unrolling models [10], [23]. Additionally, emerging Transformer-based frameworks have shown promise in capturing long-range spectral dependencies, further enhancing fusion performance [26], [27]. However, the process of unsupervised fusion refers to dense prediction problems, where each pixel in the output must be estimated without direct supervision. Additionally, most existing approaches attempt to enlarge the receptive field by adding more layers or using larger convolution kernels, which significantly increases the number of parameters and computational burden. This improvement in model understanding comes at the cost of reduced stability and efficiency [9], [12], [24]. Overcoming these limitations continues to be a focus in developing more effective fusion algorithms.

To address the above limitations, we propose a dilated transformation-guided unsupervised multimodal learning (DTUML) approach to fuse LR-HSIs and HR-MSIs, enhancing the network's representation ability without increasing additional parameter overhead. The main contributions of DTUML can be summarized as follows.

1) The theory of dilated convolution is utilized to expand the receptive field without increasing parameter overhead, thereby enabling more accurate reconstruction of spatial and spectral information.
2) Our new approach mitigates the limitation of receptive fields where encountered in conventional convolutional layers, opening a new branch for enhancing the deep model's generalization.
3) By the guidance of dilated Transformation, the hierarchical network can be more flexible to extract multiscale information to improve fusion performances.

The remainder of this article is organized as follows. Section II briefly reviews related works, referring to unsupervised HSI–MSI fusion and dilated convolution. Section III describes our DTUML in detail. In Section IV, we evaluate the effectiveness and competitiveness of DTUML using real remote sensing datasets. Finally, Section V concludes the proposed DTUML and discusses directions for future work.

II. Relevant Works

A. Unsupervised HSI–MSI Fusion

The mathematical framework of linear spectral unmixing has promoted the advancement of unsupervised HSI–MSI fusion [28]. From a data quality perspective, unsupervised HSI–MSI fusion enhances the spatial resolution of the original HSI and can be regarded as an effective strategy for achieving HSI super-resolution [18]. Several studies have leveraged this framework to address the challenges of fusing hyperspectral and multispectral data. For instance, Kawakami et al. [29] combined spectral unmixing with coupled tensor factorization to achieve an unsupervised method for HSI–MSI fusion. Similarly, Yokoya et al. [30] utilized nonnegative matrix factorization to fuse HR-MSIs and LR-HSIs to generate an HR-HSI. Wycoff et al. [31] proposed a matrix factorization-based HSI–MSI fusion approach. Yi et al. [32] integrated spectral unmixing with spatially sparse constraints to enhance performance in retaining spatial details. Despite their effectiveness, these methods primarily capture shallow spatial features, which limit their ability to uncover intrinsic pixel relationships and restrict their practical applicability in complex scenarios.

The advancement of DL has further accelerated the development of HSI–MSI fusion and hyperspectral image super-resolution [33]. DL-based unsupervised methods have emerged as a compelling alternative to traditional approaches, offering enhanced adaptability and robust performance without relying on paired training data [18], [21], [28]. These methods often employ generative models, such as autoencoders, to learn latent representations that effectively capture both spatial and spectral characteristics of hyperspectral data [34], [35], [36]. For example, some approaches embed spectral unmixing principles within autoencoder architectures, enabling HR-HSI reconstruction while maintaining physical interpretability [18], [28]. Others leverage adversarial frameworks, such as generative adversarial networks, to generate HR-HSIs by aligning spatial and spectral distributions [37]. Additionally, self-supervised learning techniques have been explored, allowing models to generate pseudo-labels or utilize the intrinsic structure of hyperspectral data for training [38]. These methods further enhance feature extraction and fusion by incorporating advanced attention mechanisms, such as spatial and channel attention, resulting in more precise and robust super-resolution across diverse scenarios [24], [28].

However, these fusion methods often rely on traditional convolutional layers to capture spatial relationships between pixels [9]. Traditional convolutions typically expand the receptive field using pooling operations for downsampling, which can reduce resolution and result in the loss of spatial details [39]. In contrast, the dilated convolutions employed in our DCDA expand the receptive field without downsampling, effectively preserving spatial information. Furthermore, traditional convolutional layers often require stacking additional layers to achieve a larger receptive field, which increases the parameter count and risks model redundancy and overfitting. Dilated convolutions address these challenges by leveraging sparse connections and improving parameter efficiency while maintaining robust performance.

B. Dilated Convolution

Dilated transformation offers a more efficient means to expand the receptive field of a network without increasing depth or filter size [40]. By introducing dilation rates, the receptive field is expanded without increasing the number of parameters or computational overhead, enabling the model to capture both local and global contextual information [41].

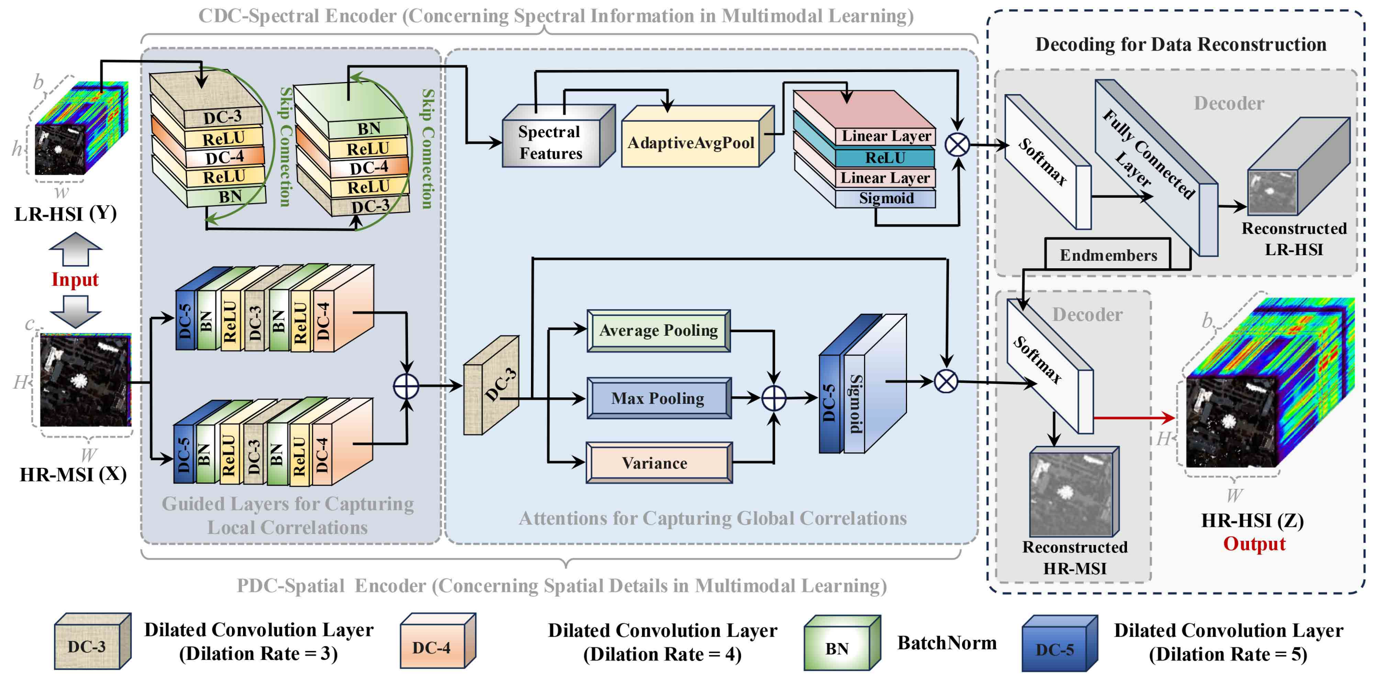

CDC-Spectral Encoder (Concerning Spectral Information in Multimodal Learning)
Decoding for Data Reconstruction
b
h
w
LR-HSI (Y)
Input
c
H
W
HR-MSI (X)
DC-3
ReLU
DC-4
ReLU
BN
Skip Connection
BN
ReLU
DC-4
ReLU
DC-3
Skip Connection
Spectral Features
AdaptiveAvgPool
Linear Layer
ReLU
Linear Layer
Sigmoid
Decoder
Softmax
Fully Connected Layer
Reconstructed LR-HSI
Endmembers
DC-5
BN
ReLU
DC-3
BN
ReLU
DC-4
DC-3
Average Pooling
Max Pooling
Variance
DC-5
Sigmoid
Decoder
Softmax
Reconstructed HR-MSI
HR-HSI (Z)
Output
Guided Layers for Capturing Local Correlations
Attentions for Capturing Global Correlations
PDC-Spatial Encoder (Concerning Spatial Details in Multimodal Learning)
DC-3 Dilated Convolution Layer (Dilation Rate = 3)
DC-4 Dilated Convolution Layer (Dilation Rate = 4)
BN BatchNorm
DC-5 Dilated Convolution Layer (Dilation Rate = 5)

Fig. 1. Flowchart of DTUML. The CDCs guide the channel attention to compose the CDC-spectral encoder that aggregates spectral correlations from the spectral domain. The PDCs and the spatial attention constitute the PDC-spatial encoder to capture spatial correlations between pixels. Here, H, h, W, and w define the spatial sizes of images, while b and c denote the spectral dimensions of ones.

Unlike traditional convolutional layers, which rely on pooling operations to enlarge the receptive field and often lose spatial details, dilated convolution preserves fine-grained details while maintaining high-resolution outputs [42]. This characteristic is especially beneficial for image super-resolution, where preserving spatial details and reconstructing high-resolution features are critical [43]. By capturing multiscale information through adjustable dilation rates, dilated convolution enhances the ability to model complex spatial dependencies, resulting in more accurate and robust super-resolution results [44]. Additionally, its sparse connection design enhances parameter efficiency, reducing the risk of overfitting while maintaining high performance, making it a powerful tool in image super-resolution tasks [45].

Although dilated convolution has been widely applied in image and signal processing, most of these methods are supervised, relying on large-scale paired datasets for training. This dependency limits their generalizability to scenarios where labeled data are scarce or unavailable. Moreover, within remote sensing, studies utilizing dilated convolutions to achieve unsupervised HSI–MSI fusion are still scarce, leaving significant room for further exploration. The dilated transformation-guided unsupervised multimodel learning framework can address key challenges such as preserving fine spatial-spectral details, capturing multiscale features, and enhancing the adaptability of models to diverse data distributions. Such new methods could bridge current research gaps and significantly advance the applications of remote sensing.

C. Channel and Spatial Attention Mechanisms

The attention mechanism is a neural network module that selectively focuses on the most informative parts of input data [46], [47]. It assigns different weights to different regions, channels, or features based on their relevance to the task, allowing the model to emphasize critical information while suppressing less important details [48]. Common forms of attention include spatial attention, which highlights significant regions in an image, and channel attention, which identifies important feature channels or spectral bands [49]. For hyperspectral and MSI fusion, attention mechanisms play a crucial role in addressing the challenges of spectral and spatial inconsistency. By applying spatial attention, the deep model can focus on high-resolution spatial details from the MSI. In contrast, spectral or channel attention enables the preservation of the rich spectral signatures in HSIs. This selective feature enhancement enables more accurate and balanced fusion results while maintaining spatial sharpness and spectral integrity. As a result, attention-based methods have become increasingly popular in remote sensing applications that demand high-quality fused imagery.

III. Proposed Method

The proposed DTUML adopts a dual-stream autoencoder architecture for unsupervised multimodal learning, as illustrated in Fig. 1. Specifically, the deep network contains the cascaded dilated convolution (CDC)-spectral encoder, parallel dilated convolution (PDC)-spatial encoder, and two decoders. The CDC-Spectral encoder is established by a CDC to guide the channel attention, while the PDC-Spatial encoder is built by a PDC to guide the spatial attention. The CDCB and PDCB are employed to aggregate local correlations of hyperspectral and MSIs, respectively. The channel and spatial attention mechanisms can capture global correlations from hyperspectral and MSIs, respectively. The CDC-Spectral and

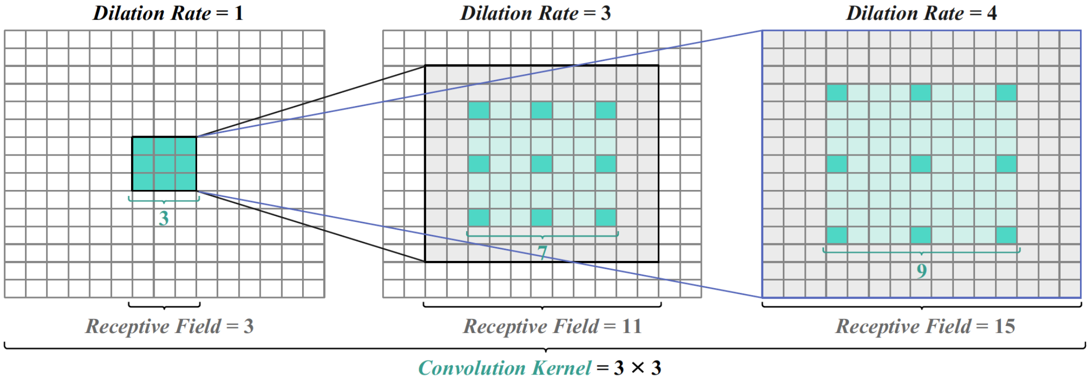


Fig. 2. Illustrating dilated convolution increases the receptive field using the different dilation rates, where the calculations of the receptive fields are carried out using (1) and (2).

PDC-Spatial encoders can implement feature embedding for spectral and spatial information. The decoders are designed to reconstruct an HR-HSI. In the MSI–HSI fusion framework, the HR-MSI is responsible for delivering spatial information, while the LR-HSI provides spectral information. To reduce the loss of spatial details, a parallel convolution block is applied in the spatial stream. In contrast, since the spectral stream focuses solely on spectral information, a cascaded convolution block is employed to better facilitate spectral feature embedding.

A. CDC-Spectral Encoder

The CDC-spectral encoder adopts CDCs to guide the channel attention for concerning spectral information in multimodal learning. The CDCs use different dilation rates to capture multiscale features hierarchically. Additionally, the CDCs adopt skip connections to allow earlier features to be concatenated later in the block. This can preserve signal integrity, especially when a signal passes through a sequence of transformations that might degrade or distort it. Skip connections can prevent the loss of signals by preserving and reintroducing the original feature information into later stages.

The CDCs set dilation rates to 3 and 4, while their convolution kernels are set to 3 × 3. The kernel has "holes" inserted between weights, allowing the convolution to cover a larger receptive field without increasing the parameters. Fig. 2 illustrates a schematic demonstrating how the receptive field is expanded in dilated convolution by configuring the dilation rate. According to [40], [41], and [42], a $k \times k$ convolution with dilation can be mathematically equivalent to an enlarged convolution. Let K define the effective kernel size, and its mathematical expression is defined as

$$K = k + (\eta - 1)(k - 1) \tag{1}$$

where k denote the convolution kernel size and η is the dilation rate. RF denotes the effective receptive field of the current layer, and it can be calculated by the following process:

$$\mathrm{RF} = K + 2(\eta - 1). \tag{2}$$

As shown in Fig. 2, following (1), a standard 3 × 3 convolution can be dilated to the enlarged 7 × 7 and 9 × 9 dilated convolutions when using $\eta = 3$ and $\eta = 4$. Meanwhile, the effective receptive fields can be computed to RF = 11 and RF = 15 by (2). Fig. 2 demonstrates that dilated convolution effectively increases the receptive field while using a small kernel size.

We use floating-point operations (FLOPs) to measure computational complexity, focusing on the number of operations performed per output pixel and multiplying that by the total number of output pixels. Let the number of input and output dimensions be d_{in} and d_{out}, respectively. The FLOPs in terms of LR-HSIs and HR-MSIs can be calculated as

$$\begin{aligned} \mathrm{FLOPs}_{(\mathrm{HSI})} &= h \cdot w \cdot d_{\text{in}} \cdot d_{\text{out}} \cdot K^2 \\ \mathrm{FLOPs}_{(\mathrm{MSI})} &= H \cdot W \cdot d_{\text{in}} \cdot d_{\text{out}} \cdot K^2. \end{aligned} \tag{3}$$

It can be clearly seen from observing (3) that the dilation rate η does not affect the number of FLOPs. Actually, the dilation rate only changes where the kernel samples input values, not the number of computations performed. Therefore, the number of parameters is determined solely by the kernel size and the input and output dimensions. In the guided layers, we combine ReLU activations to enable the model to better extract and refine semantic features. Additionally, we adopt BatchNorm to stabilize the feature distribution and reduce the overfitting risk. Meanwhile, skip connections are adopted to reintroduce the original details into deeper layers, allowing information to flow across layers to improve training performance. The cascaded structure is displayed in the top left corner of Fig. 1.

The channel attention of DTUML consists of an adaptive average pooling (AdaptiveAvgPool) [50] and a stacked network. Let $\mathbf{C} \in \mathbb{R}^{hw \times d}$ represent the spectral features obtained from the CDCs, and the process of the channel attention can be written as

$$\begin{aligned} \mathbf{C}' &= \mathrm{AdaptiveAvgPool}(\mathbf{C}) \\ \mathbf{C}'' &= \mathrm{ReLU}\left(\mathrm{MLP}\left(\mathbf{C}'\right)\right) \\ \mathbf{C}''' &= \mathrm{Sigmoid}\left(\mathrm{MLP}\left(\mathbf{C}''\right)\right) \\ \widetilde{\mathbf{C}} &= \mathbf{C} \odot \mathbf{C}''' \end{aligned} \tag{4}$$

where $\odot$ represents the Hadamard product and $\widetilde{\mathbf{C}} \in \mathbb{R}^{hw \times d}$ defines the features generated by the CDC-spectral encoder.

B. PDC-Spatial Encoder

In the MSI modal, the PDCs in guided layers are similar to those of the CDCs, where dilated rates are set to 3, 4, and 5 to obtain multiscale features. This design aims to enlarge the receptive field further, as a larger receptive field is more effective in preserving spatial details. Unlike acquiring spectral information, extracting spatial information relies more heavily on capturing contextual relationships. Therefore, we add the dilated convolution layers with $\eta = 5$ in the PDCs to enhance feature embedding for spatial information. The block of PDCs is shown on the bottom left of Fig. 1.

Assuming that $\mathbf{P}_1 \in \mathbb{R}^{HW\times d}$ and $\mathbf{P}_2 \in \mathbb{R}^{HW\times d}$ are two feature matrices acquired from the two streams referring to the PDCs. We concatenate $\mathbf{P}_1$ and $\mathbf{P}_2$ as the PDC features, and the process is expressed as

$$\mathbf{P} = \mathrm{Concat}\left(\mathbf{P}_1, \mathbf{P}_2\right)\mathbf{W} \tag{5}$$

where $\mathbf{P} \in \mathbb{R}^{HW\times d}$, $\mathbf{W} \in \mathbb{R}^{2d\times d}$ is a learnable matrix used to adjust the dimension of the features, and Concat$(\cdot)$ represents the concatenation.

The spatial attention integrates average pooling and max-pooling layers. The average pooling can highlight prominent textures, while the max-pooling reflects the overall statistical feature distribution. Let $\widetilde{\mathbf{P}} \in \mathbb{R}^{HW\times d}$ be the output of the spatial attention, and its implementation can be written as

$$\begin{aligned}\mathbf{P}_{\mathrm{Max}} &= \mathrm{Pool}_{\mathrm{Max}}\left(\mathbf{P}\right)\\ \mathbf{P}_{\mathrm{Ave}} &= \mathrm{Pool}_{\mathrm{Ave}}\left(\mathbf{P}\right)\\ \widetilde{\mathbf{P}} &= \mathbf{P} \odot \left(\mathrm{Concat}\left(\mathbf{P}_{\mathrm{Max}}, \mathbf{P}_{\mathrm{Ave}}\right)\mathbf{W}'\right)\end{aligned} \tag{6}$$

where $\mathbf{W}' \in \mathbb{R}^{2d\times d}$ is a learnable weight matrix, and $\widetilde{\mathbf{P}}$ is the features obtained by the PDC-spatial encoder.

C. Unsupervised Fusion in Decoders

According to the linear spectral mixture model, a hyperspectral or MSI can be unmixed by endmember signatures and abundance fractions. Let $\mathbf{Y} \in \mathbb{R}^{hw\times b}$ be an LR-HSI, and $\mathbf{X} \in \mathbb{R}^{HW\times c}$ be an HR-MSI. Their related spectral mixtures can be represented as

$$\mathbf{Y} = \widehat{\mathbf{A}}\mathbf{E}, \quad \mathbf{X} = \mathbf{A}\widehat{\mathbf{E}} \tag{7}$$

where $\widehat{\mathbf{A}} \in \mathbb{R}^{hw\times e}$ define the endmember matrix of the LR-HSI, $\mathbf{E} \in \mathbb{R}^{e\times b}$ describes abundances, and e represents the number of endmembers. Similarly, $\mathbf{A} \in \mathbb{R}^{HW\times e}$ and $\widehat{\mathbf{E}} \in \mathbb{R}^{e\times c}$ define the acquired endmembers and abundances associated with the HR-MSI. Let $\mathbf{Z} \in \mathbb{R}^{HW\times c}$ be an HR-HSI, and it can degrade an LR-HSI and an HR-MSI, and the degenerations can be expressed as

$$\mathbf{Y} = \mathrm{PSF}\left(\mathbf{Z}\right), \quad \mathbf{X} = \mathrm{SRF}\left(\mathbf{Z}\right) \tag{8}$$

where PSF$(\cdot)$ is the point spread function, and SRF$(\cdot)$ is the spectral response function. We further expand $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{E}}$ in (7), obtaining the following formulas:

$$\begin{aligned}\mathbf{Y} &= \widetilde{\mathbf{A}}\mathbf{A}\mathbf{E} = \widetilde{\mathbf{A}}\mathbf{Z}\\ \mathbf{X} &= \mathbf{A}\mathbf{E}\widetilde{\mathbf{E}} = \mathbf{Z}\widetilde{\mathbf{E}}\end{aligned} \tag{9}$$

where $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{E}}$ are the blurring factors of the SRF and PSF, respectively. The formulars in (9) reveals a relationship between $\mathbf{AE}$ and $\mathbf{Z}$, implying that $\mathbf{Z}$ can be reconstructed from $\mathbf{AE}$.In (7), $\widehat{\mathbf{A}}$ and $\mathbf{A}$ correspond to the hidden layers, and they are estimated by optimization. After obtaining the features in (4) and (6), we employ Softmax to activate abundance fractions of LR-HSIs and HR-MSIs

$$\begin{aligned}\widehat{\mathbf{A}} &= \mathrm{Softmax}\left(\widetilde{\mathbf{C}}\right)\\ \mathbf{A} &= \mathrm{Softmax}\left(\widetilde{\mathbf{P}}\right)\end{aligned} \tag{10}$$

where Softmax ensures the abundance values are constrained in the $0 \backsim 1$. Moreover, Softmax ensures that the sum of the elements of each column vector of the abundance matrix is equal to one. This is a physical constraint of abundance. Meanwhile, $\mathbf{E}$ and $\widehat{\mathbf{E}}$ in (7) are two learnable weight matrices between the last reconstructed layer and a hidden layer.If we regard each column of $\widehat{\mathbf{A}}$ as the output of a neuron in a linear layer, and each row of $\mathbf{E}$ as the learnable weights of a neuron, then $\mathbf{Y} = \widehat{\mathbf{A}}\mathbf{E}$ can be interpreted as a feedforward pass via a fully connected (FC) layer. We use the FC layer in the decoder of the spectral stream to learn $\mathbf{E}$. The HR-HSI $\mathbf{Z}$ can be reconstructed when $\mathbf{E}$ and $\mathbf{A}$ are available, and the reconstruction can be written as

$$\mathbf{Z} = \mathbf{AE} \tag{11}$$

where $\mathbf{Z}$ is the output reconstructed HR-HSI, and it represents the final fusion data using the proposed DTUML.

D. Loss Function

Minimizing reconstruction error (RE) is fundamental in many unsupervised learning methods, as it encourages the model to preserve essential information from the input. Considering DTUML uses an autoencoder architecture, the RE is used as a base loss ℓ_{RE}

$$\ell_{\mathrm{RE}} = \alpha \left\| \mathbf{X} - \widehat{\mathbf{X}} \right\|_F^2 + \beta \left\| \mathbf{Y} - \widehat{\mathbf{Y}} \right\|_F^2 \tag{12}$$

where $\widehat{\mathbf{X}} \in \mathbb{R}^{HW\times c}$ and $\widehat{\mathbf{Y}} \in \mathbb{R}^{hw\times b}$ are the reconstructions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. In (12), $\|\cdot\|$ denotes the Frobenius norm. α and β are two balance parameters. Using the estimated matrices $\mathbf{A}$, $\mathbf{E}$, $\widehat{\mathbf{A}}$, and $\widehat{\mathbf{E}}$ obtained from the deep network, the two reconstructed data ($\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$) can be calculated via matrix multiplication.

Since each pixel in a HSI comprises a small number of pure spectral bases, the abundance matrix $\mathbf{A}$ is expected to be sparse. Kullback–Leibler (KL) divergence is employed to promote this sparsity

$$\ell_{\mathrm{KL}} = \sum_{i=1}^{HW}\sum_{j=1}^{e}\left(s\log\left(\frac{s}{a_{i,j}}\right) + (1-s)\log\left(\frac{1-s}{1-a_{i,j}}\right)\right) \tag{13}$$

where s is a sparsity hyperparameter with a small value close to zero, and we set $s = 0.0001$ in this work, and $a_{i,j}$ is an element of $\mathbf{A}$. Combining the losses in (12) and (13), the objective function of DTUML can be written as

$$\ell = \gamma\ell_{\mathrm{RE}} + \delta\ell_{\mathrm{KL}} \tag{14}$$

where γ and δ are the balance coefficients. The pseudocode of DTUML is provided in the Algorithm 1.

Algorithm 1 Pseudocode of the Proposed DTUML

```
Input: HR-MSI $\mathbf{X}$ and LR-HSI $\mathbf{Y}$
Output: HR-HSI $\mathbf{Z}$
for Epochs do
    Set dilated rates and convolutional kernels by (1) and (2).
    Integrate the dilated convolution, ReLU, and BatchNorm layers to establish blocks.
    $\mathbf{C} \Leftarrow$ Establish CDCs to address $\mathbf{X}$.
    $\widetilde{\mathbf{C}} \Leftarrow$ Use (4) to address $\mathbf{C}$.
    Set $\eta = 3 \sim 5$ to the dilated convolutions in (1) and (2).
    Adopt the dilated convolution, ReLU, and BatchNorm layers to establish blocks.
    Use (5) to concatenate the features for PDCs.
    $\mathbf{P} \Leftarrow$ Use the PDCs to address $\mathbf{Y}$.
    $\widetilde{\mathbf{P}} \Leftarrow$ Use (6) to process $\mathbf{P}$.
    $\mathbf{E}$ and $\widehat{\mathbf{E}} \Leftarrow$ Use (7) to obtain weights.
    $\mathbf{A}$ and $\widehat{\mathbf{A}} \Leftarrow$ Use (10) to update nodes.
    $\mathbf{Z} \Leftarrow$ Use (11) to reconstruct data.
    Calculate the objective function $\ell$ by (12), (13), and (14)
    Minimize $\ell$ and update all weights.
end
```

IV. Experimental Results and Analysis

The proposed DTUML is implemented using Python 3.10.9, PyTorch 2.1.1, and CUDA 11.8. The workstation carries an NVIDIA 4080 Super GPU. Since the exact number of land-cover types present in the image is unknown, a larger number of endmembers is used to avoid omissions and ensure the model's stability. Therefore, this work sets the number of endmembers to 120. In model training, we use Adam [51] as the optimizer and set the learning rate to 3×10^{-3}, with the batch size being 64.

A. Evaluation Metrics

This study adopts several metric quantitative analyses to evaluate model performance objectively. Specifically, we evaluate the effectiveness of DTUML from different perspectives to reduce individual differences and subjective bias. The objective evaluation metrics mainly include the memory consumption (Para) over 10 000 iterations and the time consumption (Times) per 1000 iterations, as well as the following six evaluation indicators: peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), root mean square error (RMSE), structural similarity index measure (SSIM), efficiency relative to a Gaussian signal (ERGAS), and universal image quality index (UIQI). In the following quantitative analysis, an upward arrow "↑" indicates that a higher value reflects better performance, whereas a downward arrow "↓" signifies that a lower value indicates better performance.

The PSNR represents the ratio between the maximum possible power of a data signal and the power of the RE. It is commonly used to evaluate the reconstruction quality of data across different spectral bands [52]. The SAM is commonly used to calculate the spectral angle between the resulting image and the reference image, quantifying the amount of spectral information preserved in individual pixels [53]. The RMSE measures the degree of difference between the reference image and the fused image. This metric is widely used to evaluate the spectral quality of fused images. The SSIM measures the similarity between two images from three aspects: luminance, contrast, and structural information [54]. The SSIM value typically ranges from 0 to 1, where 1 indicates perfect similarity, 0 indicates no linear correlation. The ERGAS measures the overall spectral distortion between a processed (e.g., fused or reconstructed) image and a reference image [55]. A lower ERGAS value indicates better quality and closer similarity to the reference image. UIQI is a general-purpose image quality metric that models image distortion based on human visual perception [56]. The UIQI value is also usually the range of 0–1, and it is similar to the SSIM.

B. Datasets

This article mainly evaluates the fusion performance of DTUML using five real remote sensing datasets. These datasets were published by different countries, including the Chikusei,[1] Houston University 2018,[2] Washington DC,[3] Tiangong-1,[4] and ZY-1 02D[5] data. The Chikusei, Houston, Washington DC, and Tiangong-1 datasets are common datasets for testing MSI–HSI fusion performance. The four datasets use the original HSI as the GT, and LR-HSIs and HR-MSIs were obtained by scale degradation. This allows us to compare the fused HR-HSI with the GT, enabling the quantitative evaluation of the data fusion quality. The ZY-1 02D satellite is equipped with a hyperspectral camera and a multispectral camera, thereby providing two full-resolution images: the ZY-1 02D MSI and ZY-1 02D HSI. However, it is worth noting that the HSI and MSI are provided by two distinct physical cameras, and there is no GT for assessing the data quality. Therefore, a quantitative comparison of the fusion results is not possible, and only a qualitative comparison can be implemented as a supplement to other data experiments.

The Chikusei dataset was acquired by the Headwall hyperspectral sensor (VNIR-C) over Chikusei City, Japan [57]. The original image has a spatial size of 2517×2335 pixels and covers a spectral range of 363∼1018 nm. To facilitate the experiments, we extracted a subimage, and the size is $400 \times 400 \times 110$. The RGB composition is as follows: red (28th band), green (18th band), and blue (5th band).

The Houston 2018 dataset was released by the Hyperspectral Image Analysis Laboratory for the 2018 IEEE GRSS Data Fusion Contest [58]. Captured using the CASI-1500 sensor, the dataset consists of 48 spectral bands with a spatial resolution of 1 m, covering the spectral range of 380–1050 nm. After removing noisy and water absorption bands, we selected a subimage of size $400 \times 400 \times 46$ for our experiments. Its

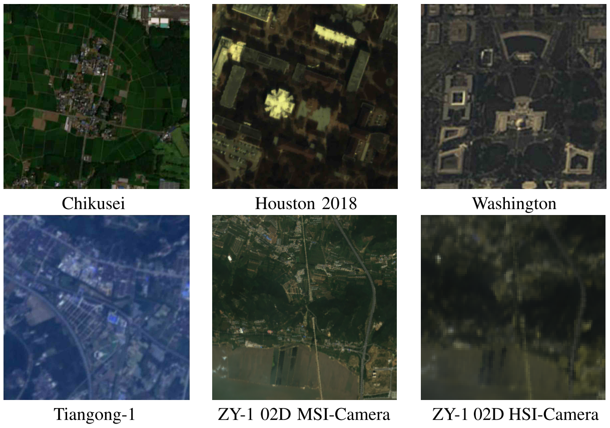
[1] http://naotoyokoya.com/Download.html

[2] https://github.com/YuxiangZhang-BIT/Data-CSHSI

[3] https://engineering.purdue.edu/biehl/MultiSpec/hyperspectral.html

[4] https://analysis.msadc.cn/org-portal/CMSDESP/dataset

[5] https://drive.google.com/drive/folders/1JLCCB6ld5R49HDLN5SsMISx1d0fuqRjO



Chikusei Houston 2018 Washington

Tiangong-1 ZY-1 02D MSI-Camera ZY-1 02D HSI-Camera

Fig. 3. Hyperspectral remote sensing images were captured by different sensors, with each scene filmed at different times and locations.

RGB composition is red (58th band), green (34th band), and blue (19th band).

The Washington DC dataset was captured by the HYDICE airborne sensor in 1995. It contains 210 spectral bands spanning the 400–2500 nm, with the spatial resolution of 2.5 m, and the image size of 1280 × 307 pixels. In our experiments, we used 191 bands after excluding 19 noisy bands. The RGB composition of the Washington DC dataset is red (55th band), green (25th band), and blue (15th band).

The Tiangong-1 dataset was collected by the hyperspectral imager onboard the Tiangong-1 satellite [59]. A representative subimage of size 240 × 240 × 54, covering the visible to near-infrared spectrum, is used as the GT. The Tiangong-1 data includes 64 effective spectral bands across the visible, near-infrared, and shortwave infrared regions, spanning the 0.4–1 μm range, with a spatial resolution of 10 m and a spectral resolution of 10 nm. The RGB composition is red (29th band), green (19th band), and blue (6th band).

The ZY-1 02D dataset was obtained from the China–Brazil Earth Resources Satellite (CBERS) program. The satellite is equipped with both hyperspectral and multispectral cameras, enabling it to capture hyperspectral and MSIs of the same scene. The CBERS series is widely known in China as "ZY-1." The hyperspectral camera integrates 166 bands (76 VNIR, 90 SWIR) within 400–2500 nm and offers a spatial resolution of 30 m with a 60 km swath [16]. The multispectral camera covers 8 spectral bands (452–1047 nm) and achieves a spatial resolution of 10 m and a swath width of 115 km to acquire HR-MSIs, both superior to its hyperspectral counterpart [60]. The HSI and MSI data from the ZY-1 02D satellite maintain their original pixel dimensions and radiometric quality. Since they have not been downsampled or compressed from the original acquisition format, they are considered full-resolution images.

The above datasets are shown in Fig 3. Table I lists the spectral ranges and sizes of all the datasets used in this article.

TABLE I
INFORMATION OF EXPERIMENTAL DATASETS, INCLUDING SPECTRAL RANGES (NANOMETER) AND DATA SIZES

Data	Spectral Range	HR-HSI Size	LR-HSI Size	HR-MSI Size
Chikusei	363-1018	400 ×400 × 110	25 ×25 × 110	400 ×400 × 8
Houston 2018	380-1050	400 ×400 × 46	40 ×40 × 46	400 ×400 × 8
TianGong-1	434-2442	240 ×240 × 56	20 ×20 × 54	240 ×240 × 8
Washington DC	400-2500	300 ×300 × 191	30 ×30 × 191	300 ×300 × 8
ZY-1 02D MSI	452-1047	/	/	300×300 × 8
ZY-1 02D HSI	400-2500	/	100 ×100 × 166	/

TABLE II
EFFECTIVENESS EVALUATION FOR THE MODULES OF THE DILATED CONVOLUTION BLOCKS ACROSS THE FOUR REAL DATASETS. "WITHOUT SCS" INDICATES THAT SKIP CONNECTIONS ARE NOT USED, WHILE "WITHOUT DCS" MEANS THAT DILATED CONVOLUTION-BASED GUIDED LAYERS ARE NOT EMPLOYED

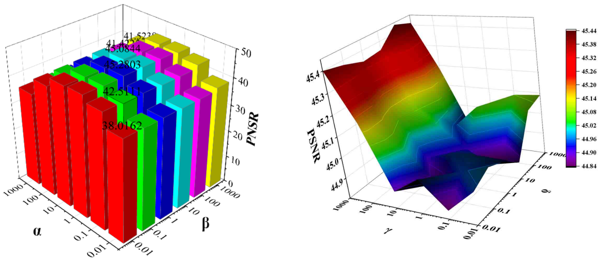
Datasets	Methods	SAM↓	PSNR↑	ERGAS↓	RMSE↓	SSIM↑
Chikusei	Without SCs	0.6954	53.6447	0.5497	0.0027	0.9942
	Without DCs	0.6925	53.6837	0.5435	0.0027	0.9943
	DTUML	**0.6889**	**53.7215**	**0.5479**	**0.0027**	**0.9945**
Houston	Without SCs	0.8898	51.3659	0.4601	0.0028	0.9960
	Without DCs	0.8936	50.7834	0.4728	0.0028	0.9959
	DTUML	**0.8525**	**52.0920**	**0.4501**	**0.0027**	**0.9961**
Tiangong-1	Without SCs	0.9281	44.7588	0.1860	0.0052	0.9871
	Without DCs	0.9197	45.0147	0.1845	0.0052	0.9870
	DTUML	**0.9139**	**45.0844**	**0.1815**	**0.0051**	**0.9873**
Washington D.C	Without SCs	1.8842	43.1526	1.1824	0.0092	0.9798
	Without DCs	1.9678	43.0970	1.1824	0.0099	0.9793
	DTUML	**1.4028**	**44.0473**	**0.8842**	**0.0065**	**0.9873**

C. Ablation Experiments

Table II presents the quantitative comparisons for the ablation studies on skip connections and dilated convolutions across different datasets. By evaluating the results using metrics such as the SAM, PSNR, ERGAS, RMSE, and SSIM, we observe that incorporating skip connections consistently improves model performance. In Table II, "without SCs" indicates that the proposed approach does not employ skip connections, while "without DCs" denotes that the proposed method removes dilated convolutions. The dilated convolution blocks, designed with a multipath structure, enhance the overall model capability by effectively enlarging the receptive field of each convolution kernel. This enables the network to capture richer spatial correlation features and more effectively preserve spatial information in the HR-MSI. Furthermore, the results in Table II demonstrate that the use of skip connections not only improves fusion performance but also enhances the stability of the deep model.

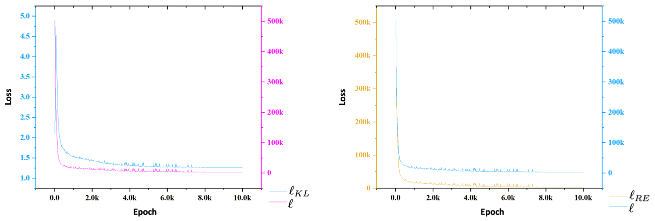
Considering the loss function comprises four balance parameters, the components constrained by these parameters all have a certain impact on model training. To investigate the influence of losses, we conduct a series of ablation studies. Fig. 4 illustrates the impact of varying parameters (α, β, γ, and δ) on the performance of the fusion process, utilizing the TianGong-1 dataset to conduct the ablation experiment. In Fig. 4(a), we can observe that with γ and δ fixed at 1, the fusion achieves the optimal performance when α and β are also set to 1. Fig. 4(b) demonstrates that the best fusion performance is achieved when γ and δ are set to $\gamma = 1000$ and $\delta = 100$, where $\alpha = \beta = 1$. In the subsequent experiments, we empirically set $\gamma = 1$, $\delta = 1$, $\gamma = 1000$ and $\delta = 100$.

Additionally, to further assess the influence of the RE and KL components in the loss function on model training, we



(a) (b)

Fig. 4. Assessment quality of data fusion under parameter settings, regarding α, β, γ, and δ. (a) Fusion accuracy with different α and β. (b) Accuracy comparison for setting different γ and δ values.



(a) (b)

Fig. 5. (a) Relationship between reconstruction loss ℓ_{RE} and total loss ℓ. (b) Relationship between KL loss ℓ_{KL} and total loss ℓ.

evaluated their respective losses separately. When any of γ and δ is set to 0, the corresponding term in the loss function becomes inactive and no longer contributes to the optimization. The loss function will remain the KL term if $\gamma = 0$, while the loss function only keeps the RE component if $\delta = 0$. Fig. 5 gives the comparison of convergence behaviors. In Fig. 5(a), ℓ_{RE} means that the loss function only adopts the RE loss. In contrast, Fig. 5(b) compares the convergences using the KL loss and total loss. According to Fig. 5, we can see that both the RE and KL terms in the loss function contribute positively to model training, thereby accelerating convergence.

The number of endmembers also affects the fusion performance, making it a hyperparameter that requires user specification. To evaluate the impact of endmember quantity on data fusion, we configure DTUML with varying numbers of endmembers and conduct experiments on four datasets: Chikusei, Houston, Washington DC, and TianGong-1, as shown in Fig. 6. The PSNR and SAM values are presented for the Chikusei, Houston, Washington DC, and TianGong-1 datasets under varying endmember numbers: 20, 40, 60, 80, 100, 120, 140, 160, and 180. It is clear that as the number of endmembers increases from 20 to 120, both PSNR and SAM values improve progressively. However, when the number increases beyond 120 to 180, performance begins to degrade. This indicates that setting the number of endmembers to 120 yields the best performance. To ensure comprehensive coverage of land-cover categories, the number of endmembers is consistently set to 120 in the subsequent experiments.

D. Baselines

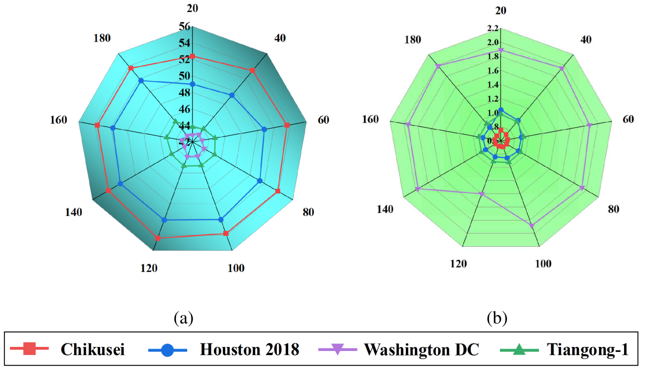
This article selected a series of representative baseline HSI–MSI fusion algorithms for comparison, including ten representative methods. These methods comprise four traditional method: CNMF [30] and CSTF [61]; as well as five DL-based methods: HyCoNet [36], UDALN [28], EU2ALN [62], CUCaNet [35], MIAE [63], uHNTC [26], and CYformer [64].

In all experiments, we adopted a rigorously controlled setup to ensure the reproducibility and fairness of the results. Through theoretical analysis and quantitative validation, we systematically compared the performance of the proposed DTUML approach with these baseline algorithms. To assess fusion results, real HSI and MSI image component datasets containing multiple sets of data were used in the experiments. Furthermore, we conducted multiple experiments to evaluate our model's reliability.

E. Comparative Experiments

To intuitively present the experimental results, we plotted performance comparison charts of different algorithms under various input scenarios. Additionally, four different heatmaps were used to visually illustrate the performance of each method. As shown in Figs. 7–9, the fusion results on different datasets are illustrated. In these Figures, each row presents different visualizations.

1) The first row shows RGB composite images, allowing a direct visual comparison of each method's performance in terms of spatial texture and overall color fidelity.
2) The second row gives SAM heatmaps, which measure the angular error between the reconstructed and GT spectra. Redder areas indicate higher errors, while bluer regions reflect more accurate spectral reconstruction.
3) The third row displays mean relative absolute error (MRAE) heatmaps, which represent the relative pixel-level deviation. A redder color suggests a larger difference from the GT values, whereas a bluer tone indicates closer numerical accuracy.
4) The fourth row presents the residual heatmap for the 10th spectral band, highlighting pixel-wise differences at that specific band. Dark blue denotes small residuals, and dark red signifies larger errors, further validating the modeling precision of each method on individual bands.



(a) (b)

Fig. 6. Spider charts of PSNR and SAM for different datasets under different endmember numbers. (a) PSNR values for setting different endmember numbers. (b) SAM values when using different endmember numbers.

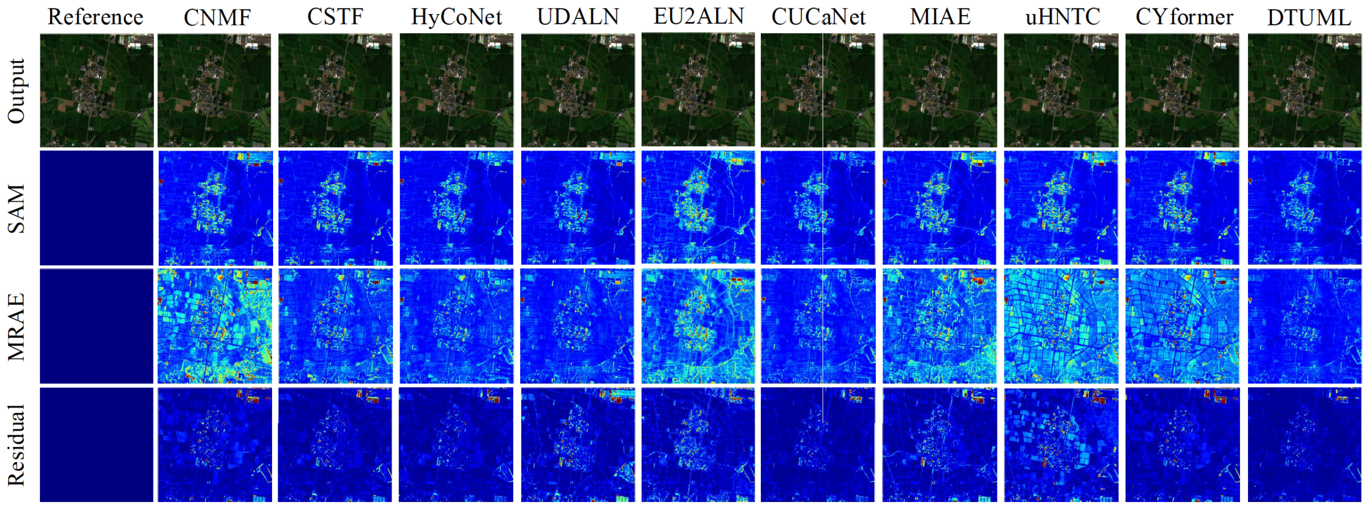


Fig. 7. Fusion results with the Chikusei dataset (first row) RGB compositions (R:58, G:38, B:20), (second row) heatmap of SAM error, (third row) heatmap of MRAE, (fourth row) residual heatmap at band 10. The error ranges of the three groups are [0, 6.5], [0, 0.15], and [0, 0.009], respectively.

TABLE III

Quality Evaluation for Different Fusion Methods, Using the Chikusei Dataset. The Best Values Are Shown in Bold

Method	SAM↓	PSNR↑	ERGAS↓	RMSE↓	SSIM↑	UIQI↑	Para (KB)	Time (S)
CNMF	0.9178	47.7283	2.2246	0.0041	0.9909	0.9976	385420	105.34
CSTF	0.1860	49.5640	2.1228	0.0037	0.9917	0.9986	384560	95.22
HyCoNet	0.7755	51.4062	1.6993	0.0032	0.9933	0.9988	478470	264.48
UDALN	0.7440	51.4933	1.1454	0.0030	0.9943	0.9991	449850	106.99
EU2ADL	1.0338	48.2605	1.2634	0.0045	0.9940	0.9984	429440	243.07
CUCaNet	0.7261	52.8972	1.4914	0.0029	0.9941	0.9990	452500	250.26
MIAE	0.9079	48.8311	0.6467	0.0041	0.9937	0.9981	449820	244.12
uHNTC	0.6918	48.3990	1.3059	0.0035	0.9927	0.9976	454570	289.81
CYformer	0.7218	49.5479	0.7295	0.0044	0.9932	0.9981	487825	292.18
DTUML	**0.6889**	**53.7215**	**0.5479**	**0.0027**	**0.9945**	**0.9992**	412100	230.80

TABLE IV

Comparison of Fusion Results on the Houston 2018 Dataset. The Best Values Are Shown in Bold

Method	SAM↓	PSNR↑	ERGAS↓	RMSE↓	SSIM↑	UIQI↑	Para (KB)	Time (S)
CNMF	0.9958	49.4645	0.5202	0.0032	0.9955	0.9995	335690	122.36
CSTF	1.3905	47.1056	0.6861	0.0044	0.9940	0.9993	332456	124.80
HyCoNet	1.6214	44.6362	0.5347	0.0053	0.9935	0.9991	355432	139.54
UDALN	0.9375	50.2141	0.4965	0.0030	0.9957	0.9996	383210	164.36
EU2ADL	2.2065	41.7955	1.2183	0.0076	0.9903	0.9982	419200	152.18
CaCuNet	0.9396	50.3476	0.4945	0.0030	0.9955	**0.9997**	367654	154.62
MIAE	0.9830	43.9933	0.6174	0.0048	0.9956	0.9991	366543	147.65
uHNTC	0.9210	49.0160	0.5720	0.0035	0.9941	0.9993	445840	330.52
CYformer	0.9511	50.0512	0.7052	0.0038	0.9945	0.9986	464130	359.63
DTUML	**0.8525**	**52.0920**	**0.4501**	**0.0027**	**0.9961**	**0.9997**	362100	152.87

TABLE V

Comparison of Fusion Results on the Washington DC Dataset

Method	SAM↓	PSNR↑	ERGAS↓	RMSE↓	SSIM↑	UIQI↑	Para (KB)	Time (S)
CNMF	2.1178	38.9928	1.3749	0.0100	0.9729	0.9983	356230	142.36
CSTF	2.1599	38.154	1.2346	0.0083	0.9543	0.9987	354850	109.44
HyCoNet	1.9628	42.5683	1.2495	0.0097	0.9800	0.9986	388200	153.20
UDALN	1.9115	40.5550	0.9589	0.0073	0.9869	0.9993	442850	204.36
EU2ADL	1.9495	42.3939	1.2484	0.0096	0.9808	0.9986	441760	174.01
CaCuNet	1.9970	41.7620	1.2658	0.0097	0.9785	0.9986	422870	188.46
MIAE	1.9646	42.1673	1.2587	0.0097	0.9794	0.9986	397200	165.23
uHNTC	2.0434	40.1107	0.8958	0.0094	0.9341	0.9966	475820	248.51
CYformer	1.9280	42.2157	0.9356	0.0070	0.9657	0.9982	472131	254.43
DTUML	**1.4028**	**43.9051**	**0.8842**	**0.0065**	**0.9875**	**0.9989**	422100	185.63

TABLE VI

Comparison of Fusion Results on the Tiangong-1 Dataset

Method	SAM↓	PSNR↑	ERGAS↓	RMSE↓	SSIM↑	UIQI↑	Para (KB)	Time (S)
CNMF	0.9811	43.5272	0.5314	0.0061	0.9858	0.9995	308540	23.51
CSTF	1.1325	41.7881	0.6219	0.0069	0.9845	0.9995	328500	30.80
HyCoNet	1.4002	38.3884	0.8810	0.0101	0.9760	0.9991	342500	35.70
UDALN	0.9924	44.0405	0.3725	**0.0051**	0.9873	0.9995	415230	50.29
EU2ADL	1.2685	39.9253	0.7493	0.0084	0.9809	0.9992	455930	50.20
CUCaNet	1.1612	41.8710	0.2090	0.0069	0.9855	0.9995	365820	42.98
MIAE	1.0967	42.6798	0.5864	0.0064	0.9846	0.9994	351420	40.53
uHNTC	0.9673	43.2571	0.5290	0.0069	0.9848	0.9993	454260	52.69
CYformer	0.9582	42.2872	0.6961	0.0062	0.9835	0.9995	479850	55.87
DTUML	**0.9139**	**45.0844**	**0.1815**	**0.0051**	**0.9873**	**0.9996**	342100	35.41

Additionally, the evaluation metrics presented in Tables III–VI, including the PSNR, SAM, RMSE, SSIM, UIQI, Parameter calculation (Para), and Computation Time. The evaluation offers a relatively objective and reliable assessment of the performance differences among various fusion methods in terms of spatial detail preservation, spectral consistency, and accuracy.

Fig. 7 indicates that the proposed DTUML demonstrates superior performance in spatial reconstruction accuracy, spectral fidelity, and numerical error control. Compared to other methods, the images produced by DTUML most closely resemble the reference images in terms of object boundaries, texture details, and overall tone. In the SAM heatmap, most regions appear in blue, with sparse and scattered red areas, indicating that DTUML effectively preserves spectral consistency. The MRAE heatmap shows that the proposed DTUML maintains low relative errors across most regions. Residual maps indicate that DTUML maintains high accuracy in modeling individual spectral bands, with only a few scattered red regions, reflecting its stability across different bands.

Table III reveals significant differences in the performance of various methods on the HSI–MSI fusion task. Specifically, a lower SAM value indicates a closer match between the reconstructed and GT spectra. Although CSTF achieved an exceptionally low SAM of 0.1860, the SAM values for other methods range between 0.7 and 1.0. In this experiment, DTUML achieved a SAM of 0.6889, reflecting commendable spectral reconstruction accuracy. DTUML achieved the highest PSNR value of 53.7215, substantially

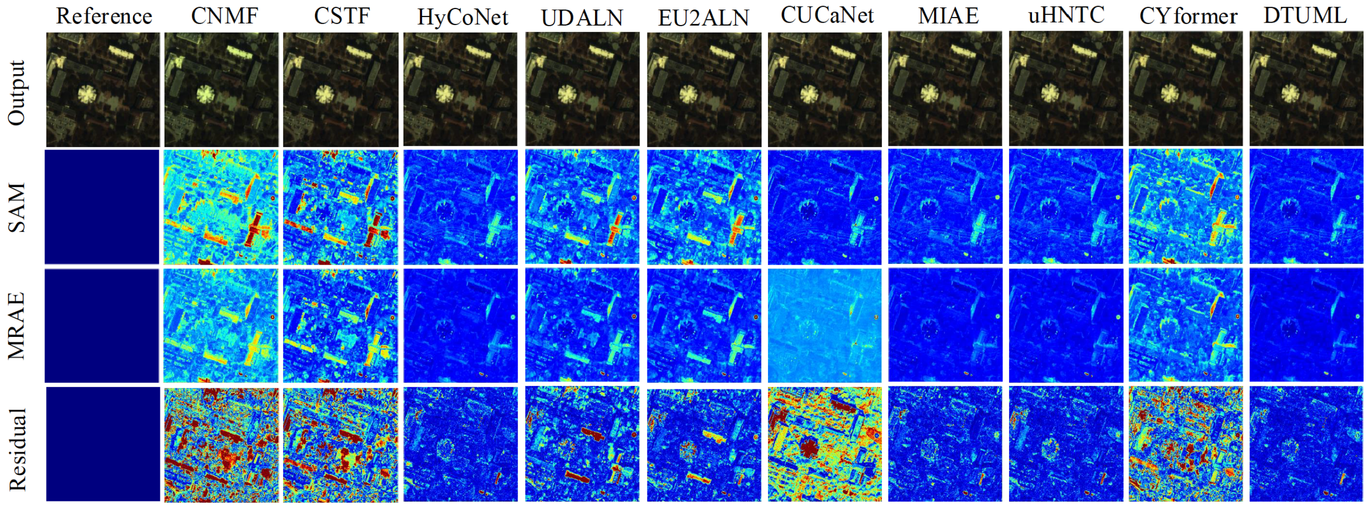


Fig. 8. Fusion results with the Houston 2018 dataset (first row) RGB compositions (R:45, G:30, B:14), (second row) heatmap of SAM error, (third row) heatmap of MRAE, (fourth row) residual heatmap at band 10. The error ranges of the three groups are [0, 6.5], [0, 0.15], and [0, 0.009], respectively.

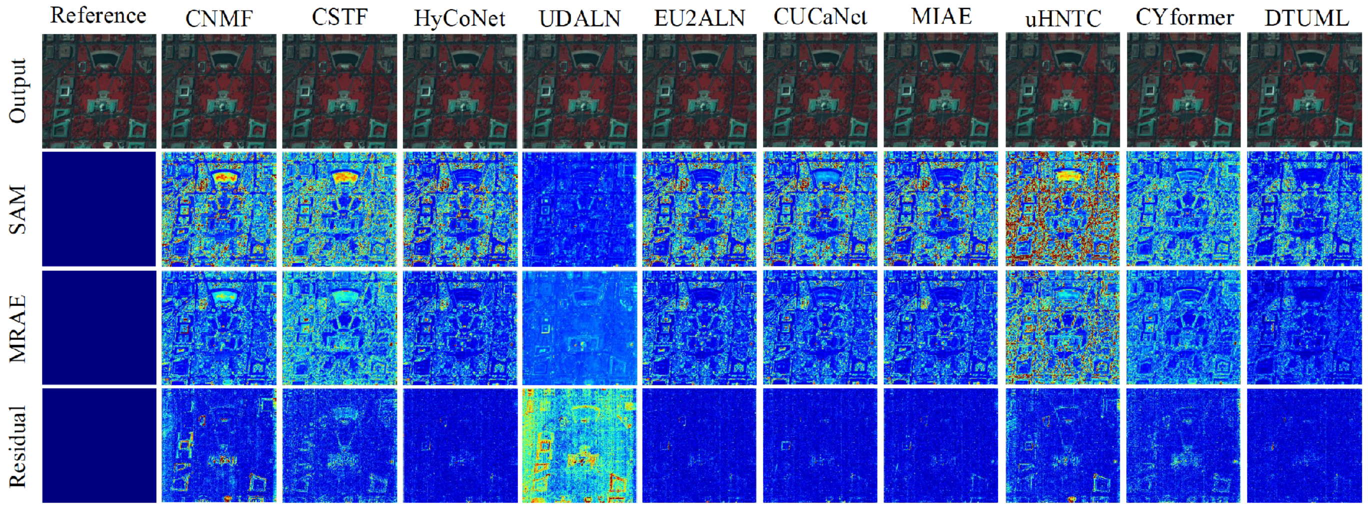


Fig. 9. Fusion results with the Tiangong-1 dataset (first row) RGB compositions (R:25, G:16, B:3), (second row) heatmap of SAM error, (third row): heatmap of MRAE. (Fourth row) Residual heatmap at band 10. The error ranges of the three groups are [0, 6.5], [0, 0.15], and [0, 0.009], respectively.

outperforming other methods, which highlights its strong capability in detail preservation and noise suppression. Additionally, the ERGAS score of DTUML is 0.5479, significantly better than EU2ADL (1.2634) and MIAE (0.6467), indicating excellent error control. The lowest RMSE value further confirms that our DTUML has less information loss than other fusion methods.

As illustrated in Fig. 8, the comparative results between the DTUML and the other fusion approaches demonstrate that DTUML achieves remarkable performance across multiple evaluation metrics. Notably, in urban environments, it effectively preserves intricate textures. DTUML consistently maintains a low SAM, even in complex architectural environments and high-contrast regions, exhibiting minimal red bias. It achieves low relative absolute error across most areas, ensuring high fidelity in pixel-level reconstruction. Furthermore, the residuals across all spectral bands remain minimal, highlighting its effectiveness in preserving the integrity of multiwavelength spectral information.

Table IV shows that DTUML achieves the best performance with a SAM value of 0.8525, indicating minimal spectral distortion. Its PSNR score of 52.0920 highlights a clear advantage in detail preservation and noise suppression, significantly outperforming other methods. Although DTUML does not achieve the top score in ERGAS, it attains the highest values in RMSR, SSIM, and UIQI. Together, Fig. 8 and Table IV further validate the effectiveness and competitiveness of DTUML.

Fig. 10 highlights the superior fusion performance of DTUML on the Washington DC dataset. A comparison of the SAM heatmaps across various algorithms reveals that DTUML's map is predominantly characterized by blue tones, with very limited red regions, indicating consistently low spectral angle errors, even in complex or high-contrast areas, and a spectral reconstruction that closely approximates the GT.

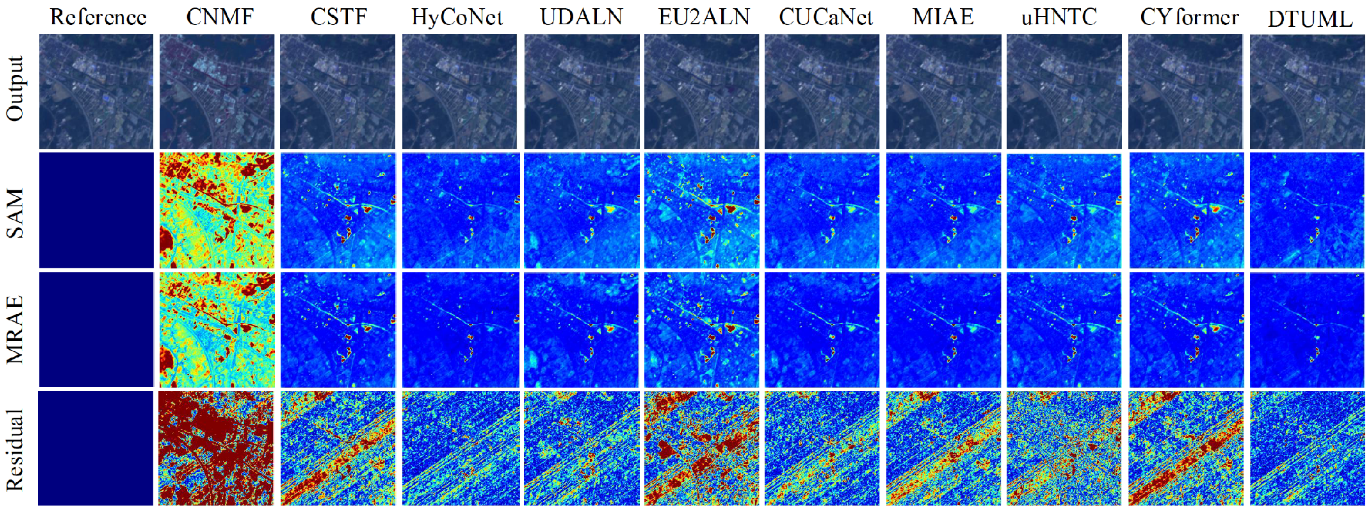


Fig. 10. Fusion results with the Washington DC dataset (first row) RGB compositions (R:50, G:30, B:16), (second row) heatmap of SAM error, (third row) Heatmap of MRAE, (fourth row) residual heatmap at band 10. The error ranges of the three groups are [0, 6.5], [0, 0.15], and [0, 0.009], respectively.

Furthermore, the MRAE heatmap demonstrates that DTUML achieves low relative errors across the vast majority of pixels. The residual maps reinforce this observation: DTUML's residuals are primarily deep blue, with only sparse red occurrences, underscoring its effectiveness in preserving spectral information. By contrast, traditional methods such as CNMF and CSTF tend to produce widespread errors in challenging scenes, while some DL-based approaches exhibit better performance.

As presented in Table V, DTUML achieves a SAM value of 1.4028, substantially lower than those of all competing methods, underscoring its superior capability in preserving spectral fidelity. Its PSNR score of 43.9051 is the highest among the compared fusion approaches, highlighting DTUML's effectiveness in restoring fine details while simultaneously suppressing noise. DTUML attains the highest scores in both SSIM (0.9875) and UIQI (0.9989), affirming its advantages in maintaining both structural and perceptual quality. While DTUML registers a marginally higher RMSE than UDALN, it consistently outperforms UDALN across all other key evaluation metrics, demonstrating a more balanced and comprehensive fusion performance. Fig. 10 and Table V further demonstrate the effectiveness and advantages of our DTUML on the Washington DC dataset.

Fig. 10 presents the comparison results of DTUML on the Tiangong-1 dataset. In the RGB composite images, DTUML accurately restores the colors and textures of various land features, including buildings, roads, and vegetation. The SAM heatmap shows that most of the imaging area appears blue, indicating that DTUML effectively preserves spectral information. In the MRAE heatmap, DTUML exhibits low relative errors, achieving high prediction accuracy across the majority of pixels and significantly outperforming other fusion methods. In the residual maps for the Tiangong-1 dataset, DTUML's results are predominantly deep blue, with only a few scattered red points, demonstrating its ability to achieve high-precision spectral reconstruction.

Table VI presents the quantitative comparison between DTUML and other competing methods. DTUML achieves a SAM value of 0.9139, lower than all other fusion approaches, indicating superior spectral accuracy. Its PSNR score of 45.0844 is also the highest among the evaluated methods, reflecting strong capabilities in detail preservation and noise suppression. Additionally, DTUML yields an ERGAS value of 0.1815, which is significantly better than those achieved by the other methods. Although DTUML and UDALN share the same RMSE value of 0.0051, DTUML consistently outperforms UDALN in all other metrics. With SSIM and UIQI scores of 0.9873 and 0.9996, respectively, DTUML demonstrates excellent structural fidelity. The results in Table VI confirm that DTUML achieves optimal or near-optimal performance across all key indicators, including SAM, PSNR, ERGAS, RMSE, SSIM, and UIQI, highlighting its effectiveness in hyperspectral image fusion. The qualitative and quantitative comparison results presented in Fig. 10 and Table VI clearly demonstrate that our new approach outperforms other methods in preserving data quality.

To provide a clearer comparison of image quality between DTUML and other methods, Fig. 11 presents the PSNR values for each spectral band across different approaches. As shown in Fig. 11, our proposed DTUML consistently achieves higher PSNR values across all spectral bands in experiments on the four datasets, demonstrating its stable and significant advantage in preserving the quality of fused data.

As is well known, DL inevitably involves some degree of information loss. Considering that dilated convolutions can alter data dimensions, we selected three representative land-cover types from the ZY-1 02D data scene to validate the fidelity of spectral information during the DTUML fusion process. These land-cover types are: mountain, water, and grass. Fig. 12 compares the spectral information of the original LR-HSI and the fused HR-HSI. The result in Fig. 12 shows that, although there is some subtle loss of spectral information, the distinctive characteristics of the *land-cover types are well

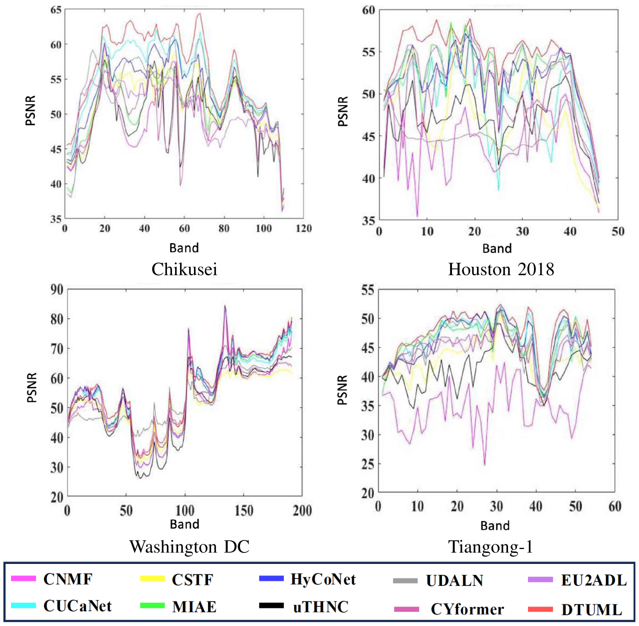


Fig. 11. PSNR values corresponding to different datasets with various bands.

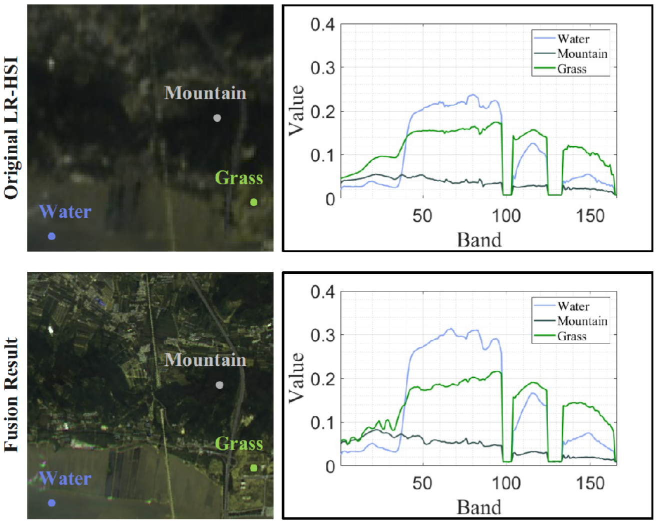


Fig. 12. Comparison of spectral curves, corresponding to the original LR-HSI and the fusion result of DTUML. (Top) Original LR-HSI. (Bottom) HR-HSI obtained by DTUML.

preserved. The spectral shape is not significantly distorted, and this data quality can substantially reduce the burden on downstream tasks.

We used our previously developed Transformer-based classification method [65] to classify the original LR-HSI, HR-MSI, and HR-HSI fused by DTUML. Since the original LR-HSI lacks a corresponding GT, we scaled the GT to facilitate selecting training samples and conducting accuracy analysis for the LR-HSI. In this experiment, we randomly selected 10% of the samples from each land-cover class in the GT for training, with the remaining samples used for testing. Fig. 13 provides the classification maps for the corresponding LR-HSI, HR-MSI, and HR-HSI. To assess classification accuracy, we used OA, AA, and Kappa as evaluation metrics.

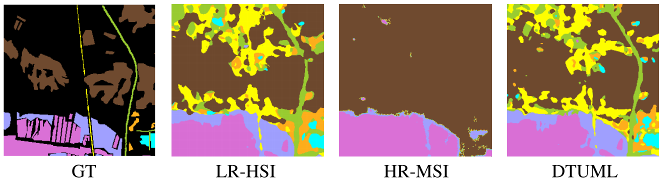


Fig. 13. Classification maps. The GT map was obtained from field surveys, and its size matches that of the original HR-MSI. Since the original LR-HSI does not have a corresponding GT, we scaled this GT to facilitate the selection of training samples and accuracy analysis for the LR-HSI.

TABLE VII

Comparison of Classification Accuracy. Using the Original LR-HSI, the Original HR-MSI, and Our Fused HR-HSI

	OA	AA	Kappa
Origial LR-HSI	91.73%	89.26%	89.51%
Origial HR-MSI	54.58%	30.17%	49.06%
Fused HR-HSI of DTUML	94.88%	93.05%	93.23%

Table VII lists the OA, AA, and Kappa coefficients for the corresponding LR-HSI, HR-MSI, and HR-HSI. From Fig. 13 and Table VII, it is evident that fusing the original LR-HSI and HR-MSI before HSI classification significantly improves classification performance. The experiments on the ZY-1 02D dataset demonstrate that our DTUML also obtains satisfactory fusion results in cross-sensor tasks.

V. Conclusion

This article proposes DTUML for fusing MSIs and HSIs. By employing dilated convolutions, the proposed DTUML effectively expands the receptive field without increasing computational overhead, enabling the model to capture both local and global spatial-spectral features. This design further ensures the preservation of spectral fidelity from LR-HSIs and spatial detail from HR-MSIs, achieving a balanced and high-quality fusion. Extensive experiments on multiple datasets demonstrate that the proposed DTUML consistently outperforms state-of-the-art fusion methods, delivering superior spatial and spectral reconstruction. The guidance of dilated transformation not only mitigates limitations such as restricted receptive fields but also enhances the model's ability to extract fine-grained multiscale information. Overall, the proposed approach offers an effective and flexible solution for HSI–MSI fusion without requiring large amounts of paired training data. Since the main focus of this work is to verify that the guidance layer built with dilated convolution can significantly enhance fusion performance, we only employ channel attention and spatial attention to capture global correlations. In future work, we plan to explore a contrastive learning strategy to further enhance the generalization capability across domains.

Acknowledgment

The authors would like to express gratitude to the Associate Editor and the Anonymous Reviewers for their constructive suggestions, which have greatly enhanced the quality of this article. Additionally, they would like to thank Dr. Haoyang Yu and Jia Jia from the Center for Hyperspectral Imaging in

Remote Sensing, Dalian Maritime University, for providing the ZY-1 02D dataset, which was used to evaluate the real-world utility of DTUML.

References

[1] Z. Li, K. Zheng, L. Gao, N. Zi, and C. Li, "Feature reconstruction guided fusion network for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 3636047.

[2] Z. Han, J. Yang, L. Gao, Z. Zeng, B. Zhang, and J. Chanussot, "Subpixel spectral variability network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5504014.

[3] H. Gao, R. Sheng, Y. Su, Z. Chen, S. Xu, and L. Gao, "Multiscale segmentation-guided fusion network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 34, pp. 6152–6167, 2025.

[4] Z. Yang, N. Zheng, and F. Wang, "DSSFN: A dual-stream self-attention fusion network for effective hyperspectral image classification," *Remote Sens.*, vol. 15, no. 15, p. 3701, Jul. 2023.

[5] Y. Su, L. Gao, A. Plaza, X. Sun, M. Jiang, and G. Yang, "SRViT: Self-supervised relation-aware vision transformer for hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 10, pp. 17585–17598, Oct. 2025.

[6] Y. Su et al., "DAAN: A deep autoencoder-based augmented network for blind multilinear hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5512715.

[7] B. Pan, Q. Qu, X. Xu, and Z. Shi, "Structure–Color preserving network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520512.

[8] H. Gao, Z. Chen, and F. Xu, "Adaptive spectral–spatial feature fusion network for hyperspectral image classification using limited training samples," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 107, Mar. 2022, Art. no. 102687.

[9] W. Chen, K. Shang, Y. Wang, W. Qi, S. Ding, and X. Zhang, "A mixed convolution and distance covariance matrix network for fine classification of corn straw cover types with fused hyperspectral and multispectral data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 134, Nov. 2024, Art. no. 104213.

[10] J. Xu et al., "FusGAT: Graph attention-based fusion network for unsupervised hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025.

[11] L. Li, H. He, N. Chen, X. Kang, and B. Wang, "SLRCNN: Integrating sparse and low-rank with a CNN denoiser for hyperspectral and multispectral image fusion," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 134, Nov. 2024, Art. no. 104227.

[12] C. Zhu, T. Zhang, Q. Wu, Y. Li, and Q. Zhong, "An implicit transformer-based fusion method for hyperspectral and multispectral remote sensing image," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 131, Jul. 2024, Art. no. 103955.

[13] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.

[14] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6503–6517, Dec. 2018.

[15] M. Xu, J. Mao, Z. Mo, X. Fu, and S. Jia, "Spectral modality-aware interactive fusion network for HSI super-resolution," in *Proc. ACCV*, 2025, pp. 301–317.

[16] H. Yu, Z. Ling, K. Zheng, L. Gao, J. Li, and J. Chanussot, "Unsupervised hyperspectral and multispectral image fusion with deep spectral–spatial collaborative constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534114.

[17] X. Feng et al., "Single space object image super resolution reconstructing using convolutional networks in wavelet transform domain," in *Proc. IEEE 3rd Int. Conf. Electron. Technol. (ICET)*, May 2020, pp. 862–866.

[18] K. Zheng, L. Gao, D. Hong, B. Zhang, and J. Chanussot, "NonRegSRNet: A nonrigid registration hyperspectral super-resolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520216.

[19] J. Zou, W. He, H. Wang, and H. Zhang, "SAM-CTMapper: Utilizing segment anything model and scale-aware mixed CNN-transformer facilitates coastal wetland hyperspectral image classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 139, May 2025, Art. no. 104469.

[20] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, Oct. 2019.

[21] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, Aug. 2022, Art. no. 102926.

[22] N. Chen et al., "Fusion of hyperspectral-multispectral images joining spatial–spectral dual-dictionary and structured sparse low-rank representation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102570.

[23] C. Zhu, R. Dai, L. Gong, L. Gao, N. Ta, and Q. Wu, "An adaptive multi-perceptual implicit sampling for hyperspectral and multispectral remote sensing image fusion," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103560.

[24] B. Tu, Q. Ren, J. Li, Z. Cao, Y. Chen, and A. Plaza, "NCGLF2: Network combining global and local features for fusion of multisource remote sensing data," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102192.

[25] J. Li, K. Zheng, L. Gao, Z. Han, Z. Li, and J. Chanussot, "Enhanced deep image prior for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5504218.

[26] X. Cao, Y. Lian, K. Wang, C. Ma, and X. Xu, "Unsupervised hybrid network of transformer and CNN for blind hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5507615.

[27] X. Cao, Y. Lian, J. Li, K. Wang, and C. Ma, "Unsupervised multi-level spatio-spectral fusion transformer for hyperspectral image super-resolution," *Opt. Laser Technol.*, vol. 176, Sep. 2024, Art. no. 111032.

[28] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[29] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *Proc. CVPR*, Jun. 2011, pp. 2329–2336.

[30] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[31] E. Wycoff, T.-H. Chan, K. Jia, W.-K. Ma, and Y. Ma, "A non-negative sparse promoting algorithm for high resolution hyperspectral imaging," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1409–1413.

[32] C. Yi, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral image super-resolution based on spatial and spectral correlation fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4165–4177, Jul. 2018.

[33] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial–spectral prior for super-resolution of hyperspectral imagery," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1082–1096, 2020.

[34] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2511–2520.

[35] J. Yao, D. Hong, J. Chanussot, D. Meng, X. X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2020, pp. 208–224.

[36] K. Zheng et al., "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2021.

[37] J. Liu, H. Zhang, J.-H. Tian, Y. Su, Y. Chen, and Y. Wang, "R2D2-GAN: Robust dual discriminator generative adversarial network for microscopy hyperspectral image super-resolution," *IEEE Trans. Med. Imag.*, vol. 43, no. 11, pp. 4064–4074, Nov. 2024.

[38] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527812.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.

[41] T. Wu, S. Tang, R. Zhang, J. Cao, and J. Li, "Tree-structured Kronecker convolutional network for semantic segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Mar. 2018, pp. 940–945.

[42] Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation," *IEEE Trans. Multimedia*, vol. 24, pp. 261–273, 2022.

[43] Z. Huang, L. Wang, G. Meng, and C. Pan, "Image super-resolution via deep dilated convolutional networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 953–957.

[44] K. Chang, M. Li, P. L. K. Ding, and B. Li, "Accurate single image super-resolution using multi-path wide-activated residual network," *Signal Process.*, vol. 172, Jul. 2020, Art. no. 107567.

[45] Z. Zhang, X. Wang, and C. Jung, "DCSR: Dilated convolutions for single image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1625–1635, Apr. 2019.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[47] L. Fang, M. Hou, B. Huang, G. Chen, and J. Yang, "DCAFusion: A novel general image fusion framework based on reference image reconstruction and dual-cross attention mechanism," *Inf. Sci.*, vol. 698, Apr. 2025, Art. no. 121772.

[48] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 1–9.

[49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–17.

[50] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8759–8768.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 8759–8768.

[52] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Quantitative quality evaluation of pansharpened imagery: Consistency versus synthesis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1247–1259, Mar. 2016.

[53] F. A. Kruse et al., "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, nos. 2–3, pp. 145–163, May 1993.

[54] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[55] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.

[56] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[57] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-5-27, 2016.

[58] B. Le Saux, N. Yokoya, R. Hansch, and S. Prasad, "2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 52–54, Mar. 2018.

[59] L. Kang et al., "Scene classification dataset using the Tiangong-1 hyperspectral remote sensing imagery and its applications," *Nat. Remote Sens. Bull.*, vol. 24, no. 9, pp. 1077–1087, 2020.

[60] J. Jia, H. Yu, C. Wang, K. Zheng, J. Li, and J. Hu, "Spectral–spatial collaborative pretraining framework with multiconstraint cooperation for hyperspectral–multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 11610–11622, 2025.

[61] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[62] L. Gao, J. Li, K. Zheng, and X. Jia, "Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509417.

[63] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522412.

[64] S. Chen, L. Zhang, and L. Zhang, "Cyclic cross-modality interaction for hyperspectral and multispectral image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 1, pp. 741–753, Jan. 2025.

[65] M. Jiang et al., "GraphGST: Graph generative structure-aware transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

Yuanchao Su (Senior Member, IEEE) received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2019.

Since 2024, he has joined the Department of Computer and Information Science, University of Macau, Macau, China, as a MYSP Post-Doctoral Fellow. From 2018 to 2019, he was a Visiting Researcher with the Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, Knoxville, TN, USA. Since 2019, he has been serving at the Department of Remote Sensing, College of Geomatics, Xi'an University of Science and Technology, Xi'an, China, where he is an Associate Professor and a Doctoral Supervisor. His current research interests include deep learning, data fusion, hyperspectral unmixing, remote sensing image processing, and graph neural networks.

Dr. Su has been a Senior Member of the IEEE Geoscience and Remote Sensing Society since 2022. He was awarded the 2023 Macau Young Scholars Program (MYSP). Moreover, he was also awarded Shaanxi Province Youth Science and Technology Star. He also serves as a reviewer and a guest editor for some international journals.

Sheng Li received the B.S. and M.Sc. degrees from Xi'an University of Science and Technology, Xi'an, China, in 2021 and 2025, respectively. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an.

His main research interests include deep learning and multimodal data fusion.

Yicong Zhou (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tufts University, Medford, MA, USA, in 2010.

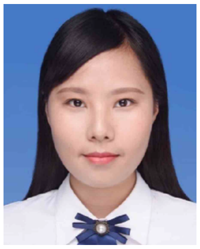
He is a Professor at the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, and artificial intelligence.

Dr. Zhou is a fellow of the Society of Photo Optical Instrumentation Engineers (SPIE) and was recognized as one of "Highly Cited Researchers" in 2020, 2021, 2023, and 2024. He serves as a Senior Area Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Lianru Gao (Senior Member, IEEE) received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 2002, and the Ph.D. degree in cartography and geographic information systems from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, in 2007.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, CAS. He is also a Visiting Scholar with the University of Extremadura, Cáceres, Spain, in 2014, and Mississippi State University (MSU), Starkville, MS, USA, in 2016. His research focuses on hyperspectral image processing and information extraction.

Dr. Gao is a Fellow of the Institution of Engineering and Technology. He was awarded the Outstanding Science and Technology Achievement Prize of the CAS in 2016 and won the Second Prize of the State Scientific and Technological Progress Award in 2018. He received the 2021 Outstanding Paper Award at the IEEE Workshop on Hyperspectral Image Processing: Evolution in Remote Sensing (WHISPERS). He is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *IET Image Processing*.



Mengying Jiang received the Ph.D. degree from the School of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, China, in 2025.

She is currently a Postdoctoral Fellow with the Department of Computer and Information Science, University of Macau, Macau, China. Her primary research interests include machine learning, graph neural networks, foundation models, bioinformatics, and hyperspectral image processing and analysis.

Xu Sun (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree in cartography and geographical information systems from the University of Chinese Academy of Sciences, Beijing, in 2011.

He is an Associate Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include hyperspectral image processing, artificial intelligence algorithms, and high-resolution remote sensing image information mining.

Haiwei Li received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Central South University, Changsha, China, in 2009, 2012, and 2016, respectively.

He is an Associate Researcher with the Key Laboratory of Spectral Imaging Technology, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China. His research interests include vicarious calibration of the hyperspectral remote sensor, hyperspectral image quality evaluation, and radiometric normalization.

Enke Hou received the B.S. and M.Sc. degrees from Xi’an Mining Institute (now renamed Xi’an University of Science and Technology), Xi’an, China, in 1984 and 1987, respectively, and the Ph.D. degree from China University of Mining and Technology, Beijing, China, in 2003.

He is a Professor and a Doctoral Supervisor at the College of Geology and Environment, Xi’an University of Science and Technology. He is the Academic Leader of the Mineral Resource Prospecting and Exploration discipline at the university and heads the scientific and technological innovation team in coal geology theory and methods. He is primarily engaged in research on coal geology and mine geology, mine water hazard prevention and control, geoscience information technology, and its geological applications.

Dr. Hou was a member of China Coal Industry Technology Committee, a member of the Coal Geology Expert Committee of China National Coal Association, a member of the Coal Geology Professional Committee of both the Geological Society of China and China Coal Society, a member of the Water Hazard Prevention and Control Professional Committee of China Coal Industry Safety Science and Technology Society, and a Standing Council Member of Shaanxi Provincial Coal Society.